**Curneu Med Tech Assignment :**

**Name :** Sahana S R

**Reg. No. :**1832043

**Dept:** MSc. Data Science 3rd year

**Diabetes Prediction**

**Aim :**

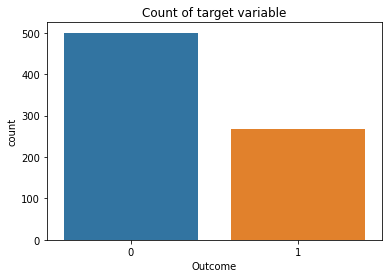
Predict which patient has diabetes from Diabetes Database.csv and try to understand the dataset attributes and try to figure out type ML model suits and build from scratch.

**Exploratory data analysis:**

**Checking Null values:**

There are no null values in the data

**Graphs:**



It can be seen from the above graph that the dataset is imbalanced



The correlation plot does not show any multicollinearity in the data

**Balancing the data:**

We use the SMOTE algorithm to balance the data. Smote stands for Synthetic Minority Oversampling Technique. SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line.

From the graph below we can see that the data is now balanced



**Train Test Split**

Of the given data 80% of the data goes to training and 20% to test set

**Scaling:**

Standard scaling technique is used to scale the data

**Model fitting:**

We fit the following models to find the best one out of it by performing hyper parameter tuning on each model. All the parameters of each model are passed into the GridSearch algorithm of python inorder to find the best parameters that produce the outcome

* Random Forest Classifier
* Logistic regression classifier
* KNN classifier
* Stochastic Gradient Descent Classifier

After fitting all the models we got the following output and hence random forest is a best possible fit for our dataset

